'Big data' approaches for novel anti-cancer drug discovery

Article (Accepted Version)

'Big data' approaches for novel anti-cancer drug discovery

Graeme Benstead-Hume, Sarah K. Wooller and Frances M. G. Pearl

Bioinformatics Group, School of Life Sciences, University of Sussex, Brighton, England

Abstract

Introduction: The development of improved cancer therapies is frequently cited as an urgent unmet medical need. Here we review how recent advances in platform technologies and the increasing availability of biological ‘big data’ are providing an unparalleled opportunity to systematically identify the key genes and pathways involved in tumorigenesis. We then discuss how these discoveries may be amenable to therapeutic interventions.

Areas covered: We discuss the current approaches that use ‘big data’ to identify cancer drivers. These approaches include genomic sequencing, pathway data, multi-platform data, identifying genetic interactions such as synthetic lethality and using cell line data. We review how big data is being used to assess the tractability of potential drug targets and how systems biology is being utilised to identify novel drug targets. We finish the review with an overview of available data repositories and tools being used at the forefront of cancer drug discovery.

Expert opinion: Targeted therapies based on the genomic events driving the tumour will eventually inform treatment protocols. However, using a tailored approach to treat all tumour patients may require developing a large repertoire of targeted drugs.

1 Introduction

Cancer represents a major and rising global health burden, with over 12 million newly diagnosed cases per annum, and is responsible for more than 15% of the world's annual deaths. The development of new improved cancer therapies is frequently cited as an urgent unmet medical need [1].

Traditionally, cancer therapies were dominated by cytotoxic agents: These therapies cause damage to DNA that exceeds a cancer cell's capacity to repair itself and have been the mainstay of cancer chemotherapy for over 30 years [2]. Although effective in treating cancers' such as testicular and breast, as well as childhood leukaemias, these agents are relatively unsuccessful in the treatment of many cancers including lung, brain, pancreatic, and oesophageal tumours, even when used in combinations [3,4].

More recently in drug discovery, the focus has changed to identifying genomic and other molecular abnormalities in cancer subtypes with a view to develop targeted therapies that offer the possibility of greater efficacy and therapeutic selectivity [5]. Early successes of single agent targeted therapies looked very promising. Two key examples, trastuzumab [6] and imatinib [7] are used in the treatment of breast cancer and chronic myeloid leukaemia respectively. Trastuzumab is a monoclonal antibody that specifically targets cells that are over-expressing the human epidermal growth factor receptor HER2/ERBB2, whilst imatinib, a small molecule, directly inhibits constitutively activated Abl kinase caused by the BRC-ABL translocation. Whilst these some of these targeted therapies have been particularly effective it appears that they may be somewhat unusual as some of the newer targeted therapies have provided at best only short-lived remissions before resistance takes hold [8,9]. However recent advances in platform technologies have provided an

unparalleled opportunity to comprehensively identify the alterations, genes, and pathways involved in tumorigenesis, raising the expectation of extending targeted therapies [10,11].

Here we review the ‘big data’ approaches for identifying the driver genes in cancer and discuss the approaches used to determine which of these would be good drug targets. We also highlight some of the excellent online data resources available to the cancer drug discovery community. Furthermore we discuss some of the challenges faced by the large repositories that contain cancer and biological data and how these issues are being addressed by new developments such as ICGCmed.

2 Cancer types and subtypes

Historically, the ~200 cancer types (and sub-types) described were characterised by the shape and location of the tumour and its growth progression. Heterogeneous populations of tumours can now be clustered into clinically and biologically meaningful subtypes using similarity of molecular profiles (e.g. [12]). As a cancer evolves, it induces dynamic changes in the genome. These include somatic mutations, copy number variations, abnormal gene expression, and deleterious epigenetic patterns. The result of this is that each individual patient's cancer will be unique. However the pathways affected in different cancer types and subtypes of tumours are similar and as such the same therapeutic strategies can often be used for groups of patients. In addition, it is these genetic and phenotypic changes that occur during tumorigenesis that alter the set of genes upon which cells become dependent. These changes generate vulnerabilities that can often be translated into successful therapeutic approaches.

2.1 Identifying cancer drivers

In the majority of cancers most of the genetic changes acquired as the disease progresses are inconsequential in terms of driving the cancer phenotype, however a few of the changes in a small set of genes are crucial to the development or the sustainability of the disease. The genes that harbour these critical alterations are called ‘driver genes’ and many studies powered by major international consortia have been undertaken to identify them (e.g. [13–16]).

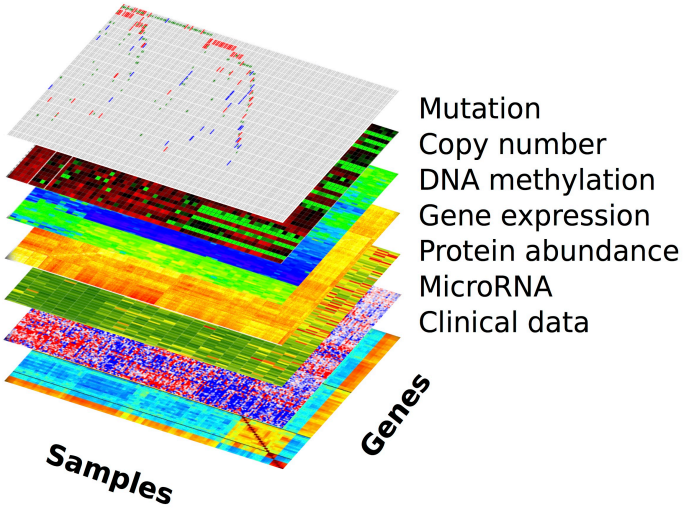
2.1.1 Sequencing approaches for identifying cancer drivers

Initial studies using ‘big data’ approaches focused on identifying mutated driver genes in groups of patient samples. The Cancer Genome Project analysed 274 megabases (Mb) of DNA, corresponding to the coding exons of 518 protein kinase genes in 210 diverse human cancers [17]. The study identified 1,000 somatic mutations and used a statistical approach to rank each of the kinases studied on the probability of it containing a driver mutation. Similarly Vogelstein and co-workers analysed the exons representing 20,857 transcripts from 18,191 genes in 11 colorectal and 11 breast tumours. Genes mutated more frequently than would be predicted by chance were identified as likely drivers. They also described the first genomic landscapes of breast and colorectal cancers [18]. As technology has improved and the costs of sequencing reduced, many other studies have worked with both disease based and pan-cancer somatic mutational data to identify ‘mutation’ driver genes (e.g. [19–21]).

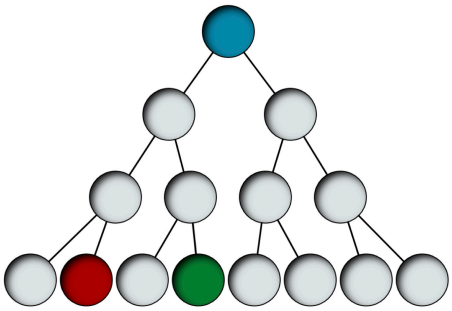
2.1.2 Multi-platform approaches for identifying cancer drivers

Due to the complexity of the changes in cancer, multi-platform approaches have also been developed. These use complementary high-throughput technologies to measure somatic mutations, as well as copy number variation (CNV), altered patterns of epigenetic modification, and changes in levels of transcription and protein expression. These large-scale studies have now surveyed ~30 individual tumour types with large cohorts of patients. Statistical approaches (e.g. [14,22]) can then be used to identify the significantly altered genes. These approaches have proven promising for identifying the highly recurrent altered genes that can provide novel drug targets (see figure 1).

Multiomics data



Model



Drug targets

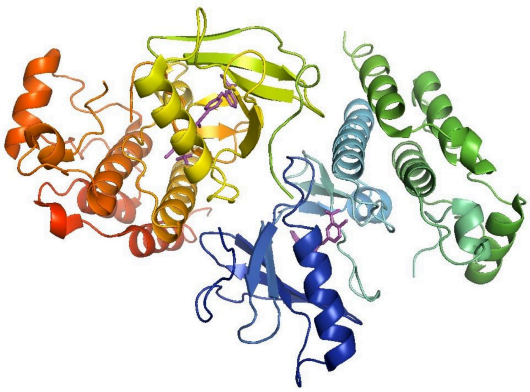


Figure 1. This figure illustrates how large-scale multi-platform analysis of large cohorts of patients can be used to identify driver genes that are potential cancer drug targets. ‘Multiomics Data’ illustrates the types of data collected for each individual data. This includes; mutation data that maps somatic genetic muta- tions in human patients that can be used to highlight genes that are recurrently mutated in certain cancer subtypes. Copy number variation maps how sections of the genome, and potentially whole genes, are repeated. Gene expression data is used to estimate transcription RNA expression levels for each gene. DNA methylation data maps epigenetic markers. MicroRNA data maps non-coding RNA molecules that function in RNA. Protein expression data maps the levels of known proteins in each cell and clinical data provides important information about the patient and their outcome. Through modelling using tools such machine learning decision tree based classifiers or other statistical methods these datasets can be analysed together to identify driver genes that have the potential to be cancer drug targets.

Brennan and co-workers [13] described the landscape of somatic genomic alterations based on the genomic and proteomic profiles of more than 500 examples of glioblastoma (GBM). In addition to alterations in the signature oncogenes of GBM, such as EGFR and PI3K, they found that over 40% of tumours contained missense mutations among the chromatin-modifier genes. Pereira et al. [14] analysed the genomic and transcriptomic landscapes of breast cancers of 2,433 breast cancers found that PIK3CA and TP53 were the most frequently mutated genes with only five other genes harbouring coding mutations in at least 10% of the samples (AHNAK2, KMT2C, GATA3, MUC16, SYNE1). Giannakis et al. [15] performed whole-exome sequencing of 619 colorectal cancers and integrated the results with tumour immunity, pathology, and survival data identifying previously undetected recurrently mutated genes.

Pickering et al. [23] identified a number of frequently mutated drivers in oral squamous cell carcinoma (OSCC) through the comprehensive genomic analysis of gene expression, methylation, copy number and point mutations. Zheng et al. [24] collected the clinical and pathologic features, genomic alterations, DNA-methylation profiles, and RNA and proteomic signatures of 91 cases of Adrenocortical Carcinoma identifying at least 6 driver genes including TP53, ZNFR3, CTNNB1, PRKAR1A, CCNE1, and TERF2.

Several studies have surveyed multi-omic data from multiple tumour types to identify driver genes (e.g. [20,25]]). For example, working at a pan-cancer level Mo et al. [26] developed iCluster+. This method performs pattern discovery that integrates diverse data types including pan-cancer somatic mutation, copy number gain, normal, loss and gene expression values to predict driving factors in cancers. These multi-

platform approaches are revealing a growing list of driver genes across the most common human cancers [16,27–29].

2.1.3 Actionable drivers

Driver genes can be broadly classified by the manner in which, when altered, they contribute to the disease process. In oncogenes an increase in activity, a gain or, more rarely, a change of function (GOF) is required for tumorigenesis whereas tumour suppressors contribute to the development of cancer when genetic changes (or epigenetic silencing) result in a loss of function (LOF) .

From a drug discovery perspective, targeting these two different types of drivers requires a very different approach. Oncogenic GOF drivers can often be targeted directly - for example dabrafenib has been approved for the treatment of late-stage melanoma, and targets the constitutively activated oncogene BRAF V600E. In lung cancer, the genetic alterations observed in EGFR (the epidermal growth factor receptor) and ALK (anaplastic lymphoma kinase) have also resulted in the development of targeted therapies. Both genes encode pharmacologically targetable tyrosine kinases involved in growth factor receptor signalling. Cetuximab, panitumumab, gefitinib and erlotinib are licensed inhibitors of the EGFR tyrosine kinase and crizotinib is an ALK inhibitor all of which are licensed for the treatment of lung cancer [30–33].

Tumour suppressors usually have to be targeted indirectly using approaches such as synthetic lethality, or by reactivation (see section 2.1.5).

2.1.3.1 The 20:20 rule

The classification of known driver genes as either tumour suppressor or oncogene is often well documented in the literature, although when a new driver is identified via a high-throughput approach, its class can often be unclear. However, the mutational patterns observed in cohorts of tumour samples differ markedly between tumour suppressor and oncogenes and several groups have used data from whole exome sequencing of large data sets to automatically distinguish between them on that basis [34–36].

Vogelstein's '20:20 rule' [16] asserts that if 20% of all mutations observed in a gene within a cohort of tumour samples are truncations, then that gene is likely to be a tumour suppressor, whereas if 20% of all missense mutations occur at a single position in the sequence, the gene is predicted to be an oncogene [37].

2.1.4 Pathway approaches for identifying actionable drivers

Even with the large quantities of patient data currently available there is not always the statistical power available to identify driver genes that occur with low frequency. Many genes within a biological pathway may contribute to tumour biology even though they are only infrequently altered. Systematic mapping of these low frequency drivers can highlight key pathways that offer 'druggable' targets for novel therapeutic strategies.

The most straightforward approach to identifying target pathways from infrequent events, is through fixed-gene set enrichment analysis. In this approach fixed gene sets are constructed based on known biological pathways and processes. It is then possible to assess whether there are more gene disruption events within the set than expected by chance (e.g. [38–43]). Known pathways can also be extended with

pathways based on genes that are predicted to have similar functionality, using tools such as GeneMania [44]. Similarly, local network properties can be used to predict which related proteins are affected and these smoothed networks are then clustered [45].

2.1.5 Synthetic lethality approaches

Driver tumour suppressors pose a particular challenge in drug discovery. Occasionally, it may be possible to directly re-activate a mutationally inactivated tumour suppressor. Post-translational re-activation of a protein requires that the inactivating mutation does not truncate or ablate the protein product but generates a form amenable to stabilisation by a modifying ligand. This approach is being explored for a subset of destabilising mutations of *TP53* (reviewed in [46]).

The alternative strategy for targeting LOF of tumour suppressors is the identification of other 'complementary' gene products that will be the actual pharmacological target using the concept of synthetic lethality. Synthetic lethal (SSL) interactions describe the situation where either gene in a pair of genes can be disrupted without significantly affecting cell viability, whilst disruptions in both genes causes cell death [47]. Synthetic sensitivity results in impaired cell growth or proliferation, which may lead to cell death in the presence of additional stresses or additional therapeutic insults. In practice, therapeutic SSL responses are likely to be on a spectrum, with the ideal being single-agent lethality.

To exploit SSL therapeutically, the genetic defects in an affected pathway must be combined with a pharmacologically induced defect in a compensating pathway [48], an approach that may provide significantly improved therapeutic indices compared to standard chemotherapies [49].

Genes involved in DNA damage response (DDR) are prime candidates for SSL interactions as there are multiple complementary pathways for repairing DNA and many of the DDR genes exhibit LOF defects in a variety of tumours [2,50]. The best current example of therapeutic exploitation of SSL is the inhibition of PARP1 [51] a key enzyme in single strand break repair which is SSL with genetic defects in the BRCA1, BRCA2 or PALB2 homologous recombination (HR) proteins commonly observed in hereditary breast, ovarian, prostate and pancreatic cancers (Figure 2).

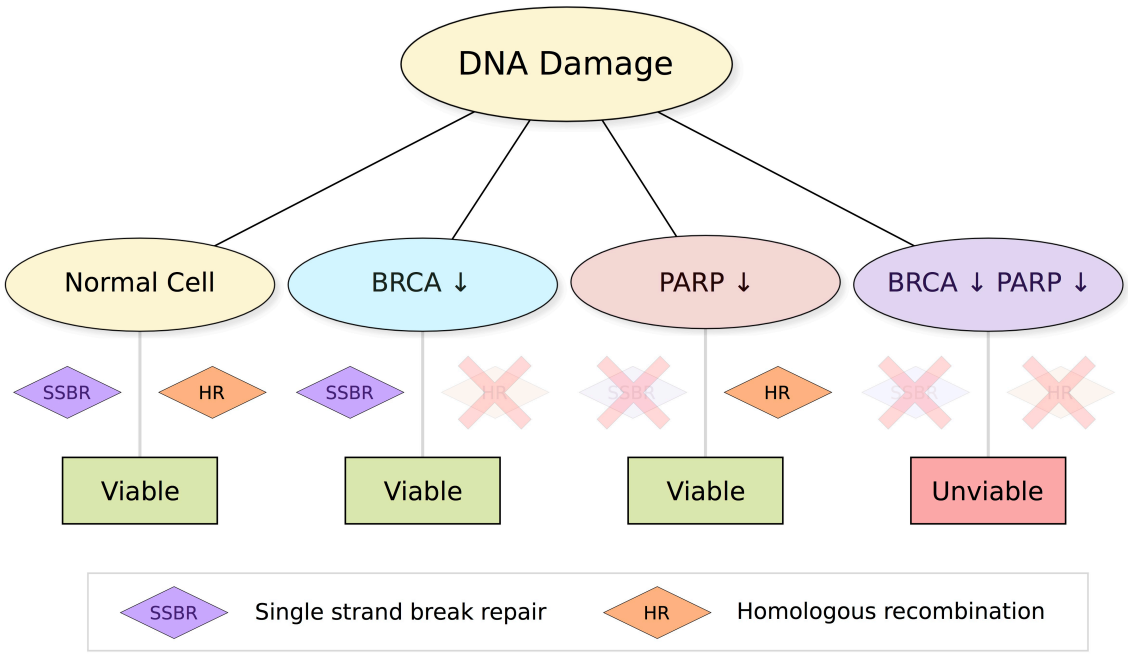


Figure 2: An illustration of the therapeutic value of synthetic lethality approaches for targeting cancer cells. When DNA is damaged in a healthy cell, it can be repaired by several mechanisms including single strand break repair (SSBR) and homologous recombination (HR). In tumour cells where BRCA1 or BRCA2 is genetically inactivated by mutation, the DNA in the cell can no longer be repaired by HR, however the cells are still viable as SSBR is still working. In healthy cells, where PARP is inactivated by targeted inhibitors, the DNA can no longer be repaired by SSBR, but the cells are still viable as HR still occurs. In tumour cells where BCRA1 or BRCA2 are genetically inactivated, and the PARP genes inactivated by PARP inhibitors damaged, DNA cannot be efficiently repaired. This leads to the accumulation of genetic damage in the cell and ultimately to cell death.

Historically a ‘hypothesis driven’ process, the prediction of SSL interactions has traditionally been based on proven associations, often related to loss of particular cell cycle checkpoints or pathways related to those of known tumour suppressors or by homology in model organisms [52,53]. This approach has worked well to identify a small numbers of SSL interactions but the experimental burden of screening many pairwise candidates had prohibited any large scale systematic studies of human synthetic lethality. The increased availability of genetically modified human cell lines and high throughput genetic screening combining RNAi screens with libraries of chemical inhibitors has allowed the search to widen [54] and a number of statistical, probabilistic and machine learning models have proved both cheaper and faster to implement compared to traditional screening methods and have demonstrated impressive levels of accuracy when predicting SSL interactions [55]. These studies have approached the prediction of SSL interactions in diverse ways - including DAISY which utilises somatic copy number and mutation profiles [56] as well as studies that focus on RNAseq data [57], phosphorylation levels between signalling pathways [58], gene ontology, co-expression data and protein interaction data [59]. Biological networks have also been a focus of study [60–62] with genome-wide protein interaction network parameter data across species being used to predict genetic interactions [63,64]. A number of collated and curated SSL interactions for humans and model organisms are available via BioGRID [59].

2.1.6 Synthetic dosage lethality (SDL)

The protein products of several potent oncogenes such as MYC and KRAS lack drug-like binding ‘pockets’ in their protein structure [65] and as a result they are mostly unamenable to current small-molecule drug discovery approaches. Big data

approaches to produce therapies for these ‘undruggable’ oncogenes include high throughput screening to find their synthetic dosage lethality partners. Similar to SSL, SDL causes cell death as a result of one gene being genetically activated (GOF, the oncogene) and another being inactivated (LOF, the drug target).

Several groups have utilised screening approaches to identify the SDL partners of activated mutant KRAS, which is frequently observed in subsets of lung and colorectal cancers [66]. A genome-wide shRNA screen in colorectal cancer cells and found that KRAS mutants were hypersensitive to APC/C and PLK1, whereas SDL partners that have been identified through RNAi screening include CDK1, CDK4, GATA2, Snail2, STK33, TBK1 and WT1 [67–72].

2.1.7 Non-oncogene addiction

The viability of a cancer cell is dependent on a variety of genes and pathways that are not inherently oncogenic themselves [37] but are essential to support the phenotype of cancer cells. This dependence - termed non-oncogene addiction - has been successfully targeted therapeutically. Examples in the clinic that use this approach include the hormone-recognition- and hormone-biosynthesis-targeting agents such as aromatase inhibitors for breast cancer and the cytochrome P450 family 17 subfamily A member 1 (CYP17A1) inhibitors for prostate cancer [29]. Where as those in the discovery phase include MTH1 which is essential in cancer cells as it cleanses oxidized dNTP pools to prevent incorporation of damaged bases during DNA replication [73,74].

One ‘big data’ approach to identifying cancer-specific vulnerabilities, including non-oncogenic addiction therapeutic targets, is to profile genetic dependencies in cancer cell lines. Campbell et al. [75] completed a large scale siRNA screening of kinase

dependencies in 117 cancer cell lines from ten cancer types. By integrating the siRNA screen data with molecular profiling data, including exome sequencing data, they demonstrated that genetic dependencies associated with specific cancer driver gene mutations could be identified. Furthermore, by identifying dependencies that were present within the known functional relationships of driver genes or kinases, they were able to predict successfully that osteosarcoma cells would be sensitive to FGFR inhibitors and SMAD4 mutant tumour cells would be sensitive to mitotic inhibitors.

2.2 Clinical trial design

As well as potentially guiding the development of therapeutics an understanding of cancer subtypes can also help provide better frameworks for clinical trial design through the stratification of patient selection. Targeted therapies are often only effective for a sub-population of patients with the specific disease, for example, whilst there may be a large number of breast cancer patients, the subset with a specific genetic mutation driving that cancer, for example a BRCA1 mutation, may be much smaller. Without factoring these disease subtypes into a trial design the efficacy rate may suffer as a result.

This notion of patient stratification has led to the development of a number of new trial frameworks such as ‘umbrella’ and ‘basket’ trials

An umbrella trial tests multiple targeted therapies based on the disease subtype. During an umbrella test a number of patients with a specific disease will be stratified based on individual mutations and assigned to different arms of the trial each of which distributes therapy specific to that disease subtype [76]. Basket trials on the other hand focus on patients with a common mutation but with varied diseases, a

BRAF mutation for example is common in a number of diseases. In a basket trial all patients will be given a the same drug that targets the shared mutation [77,78]. These trial designs provide a platform for testing several therapies concurrently for a single disease or a range of diseases with a single therapy, effectively allowing us to run a number of trials at once improving both efficiency as well as efficacy rates through patient stratification [79].

3 Target tractability

In combination, large-scale studies have uncovered extensive catalogues of genes whose protein products require assessment of their potential as viable drug targets or their 'druggablity' [28]. The druggability of a protein refers to whether the protein has the ability to bind small drug-like molecules with high affinity, this largely is dependent on its three dimensional structure

Introduced in 2002, the concept of the 'druggable genome' identified the genes within the human genome that coded for proteins that could be modulated by small drug-like proteins [80]. A recent update has mapped 1,578 FDA-approved drugs act onto 893 human drug targets [29]. The original analysis evaluated the 'druggability' of all human proteins by calculating their sequence identity to known therapeutic targets and predicted that over 10% of the human proteome was druggable. More recently methods have expanded this number of druggable proteins. Methods been developed to predict druggability on proteins whose family members have previously been untargeted analysing of the protein's 3D structure using machine learning based techniques [81–83]. Several studies have identified potential cancer drug targets from previously untargeted families using these types of approaches [2,84].

4 Cancer Data Repositories and online analysis tools

Much of the data produced from these large-scale studies have been produced by multi-national consortia and are available for independent analyses in large repositories or to browse using online tools. Here we highlight several key resources (see Table 1).

COSMIC [66], the Catalogue Of Somatic Mutations In Cancer, focuses on curating and cataloguing all known somatic mutations in human cancer. It currently describes over 4 million coding mutations from over 1.25 million tumour samples. It includes data from 24,000 published studies as well as data from 29000 whole genome samples. The data is manually curated, allowing very precise definitions of disease types and patient details.

The MOKCa database [85] has been developed to help researchers identify which genes are tumour suppressors, oncogenes and cancer drug targets and to highlight the driver mutations within them. Mutation data from the COSMIC database have been mapped to their protein products, and the mutations have been structurally and functionally annotated.

The Genomic Data Commons (GDC) [86], which already holds 4.1 petabytes of data, provides curated storage for over 14,530 cancer cases, including those that were previously curated by the The Cancer Genome Atlas (TCGA) [87,88]. The repository includes clinical and biospecimen data, and provides access to multiple ‘omic’ data types such as mRNA expression, somatic mutations, copy number variation and protein abundance. Although the GDC is of fundamental importance to researchers, at the time of writing, the transition from TCGA to GDC has not been entirely smooth: methylation data has been archived rather than brought within the GDC hub and

some data on somatic nucleotide variations as well as TCGA explanatory information is not currently available.

The International Cancer Genome Consortium (ICGC) [89] was established to coordinate large numbers of international research projects to identify the genomic changes present in a diverse range of cancers. Release 23 contains multi-platform data (somatic mutations, abnormal expression of genes, epigenetic modifications) from more than 16,000 cancer donors spanning 70 projects and 21 tumour sites. The aim of the ICGC is to make these data rapidly available to the research community, with minimal restrictions on the use of the data. ICGCmed is the next phase of the project and will link the current (and new) multi-platform data to clinical and health

Resource name	Description	Reference	URL
Cosmic	A catalogue of somatic mutations in cancer describing over 4 million coding mutations from over 1.25 million tumour samples.	[66]	cancer.sanger.ac.uk/cosmic
MOKCa	MOKCa identifies tumour suppressors, oncogenes and cancer drug targets and highlights the driver mutations within them.	[85]	strubiol.icr.ac.uk/extra/mokca
Genomic Data	GDC includes clinical and biospecimen data,	[86]	gdc.cancer.gov

Commons (GDC)	sequencing data for DNA, mRNA, and miRNA, and data on variants and mutations expression for genes, exons, and miRNA.		
The Cancer Genome Atlas (TCGA)	An original source of much of GDC's data TCGA provides similar data.	[87]	cancergenome.nih.gov
The International Cancer Genome Consortium (ICGC)	The ICGC coordinates large numbers of international research projects to identify the genomic changes present in a diverse range of cancers.	[89]	icgc.org
cBioPortal	cBioPortal allow researchers to perform integrated analysis of genetic, epigenetic, gene expression, and proteomic cancer data along with clinical profiles.	[38]	www.cbioportal.org
ChEMBL	ChEMBL is a large-scale bioactivity database containing information manually extracted from the	[92]	www.ebi.ac.uk/chembl

	medicinal chemistry literature.		
The Cancer Cell Line Encyclopedia (CCLE)	The CCLE contains pharmacological profiles for 24 anticancer drugs tested across 479 cancer cell lines.	[93]	portals.broadinstitute.org/ccle/home
The Genomics of Drug Sensitivity in Cancer (GDSC)	The GDSC contains data from over 75,000 experiments, describing the response to 138 anticancer drugs across almost 700 cancer cell lines.	[94]	www.cancerrxgene.org
NONCODE	NONCODE features non-coding RNA details and sequences for non-coding RNA in 16 species.	[95]	www.noncode.org
lncRNAdb	lncRNAdb provides non-coding RNA data including nucleotide sequences, genomic context, expression data, structural information, subcellular localisation,	[96]	www.lncrnadb.org

	conservation and functional annotation.

Table 1: Description and URL of the resources discussed in the text

information. The intention is that this should include data on lifestyle, patient history, cancer diagnosis, and response to and survival following therapies. The quality of clinical information included within current ICGC data is variable at best. ICGCmed has the potential to be game-changing in the fight against cancer, but to achieve its aims it will be important to ensure that the quality control imposed on 'omic' data is extended to clinical data.

Tools such as cBioPortal [38] allow researchers to perform integrated analysis of genetic, epigenetic, gene expression, and proteomic cancer data along with clinical profiles through an online platform for small subsets of genes (<30) [90]. It currently contains data from 147 cancer studies [38,91].

Although not containing cancer data per se, ChEMBL [92] is also a useful resource for cancer drug discovery. It is a large-scale bioactivity database containing information manually extracted from the medicinal chemistry literature. It contains information on proteins and the compounds tested against them (including their structures). It contains information on the biological and physicochemical assays performed on these targets, recorded in a structured form.

The Cancer Cell Line Encyclopedia (CCLE) [93] profiles genetic dependencies in cancer cell lines. It contains pharmacological profiles for 24 anticancer drugs tested across 479 cancer cell lines that have been characterised by gene expression, chromosomal copy number and sequencing data. Analysis of these data has allowed identification of genetic, lineage, and gene-expression-based predictors of drug sensitivity.

The Genomics of Drug Sensitivity in Cancer (GDSC) [94] database contains data from over 75,000 experiments, describing the response to 138 anticancer drugs across almost 700 cancer cell lines. Data includes somatic mutations, gene

amplification and deletion, tissue type and transcriptional data and is integrated with the cell line drug sensitivity data to provide molecular markers of drug response.
Other Databases focus on non-coding genetic material such as non-coding functional RNA, for example, NONCODE [95] which features non-coding RNA details and sequences for non-coding RNA in 16 species including humans and lncRNAdb [96] which features details on non-coding RNA including nucleotide sequences, genomic context, expression data, structural information, subcellular localisation, conservation and functional annotation with referenced literature.

5 Big data

The growth of ‘big data’ in cancer research has been intensified by the advancement of technologies such as high-throughput sequencing, the decreasing cost of these technologies, and the maturation of infrastructure in related industries. This has lead to global collaborations and an unprecedented volume of openly accessible biomedical data. The European Bioinformatics institute for example, as of December 2015, had provision to host a capacity of 75 petabytes of openly available biomedical data [97].
As a result of this growth, much of the data available via these biomedical data repositories are too large or complex to be easily processed, analysed or visualised using traditional methods. This category of data, often described as big data, presents a range of opportunities in the pharmaceutical industry, along with a number of challenges, and its use is already deeply ingrained in each step of the drug discovery pipeline.
Big data is commonly defined based on a set of characteristics such as unusually large volume, velocity and / or variety. Succinctly put big data is ‘data that is too big,

too fast or too hard for existing tools to process' [98]. However the changing nature of technology makes it hard to pinpoint exactly what constitutes big data using this definition and datasets that would have required supercomputers to manage a few years ago might now be easily processed on today's moderately powerful modern desktop. More generally datasets that require a mathematical model for their analysis rather than traditional direct analysis may fall into the category of big data [99].

5.1 Challenges of big data

While big data presents many opportunities in drug discovery it also presents challenges, both technical and conceptual. Technical challenges such as storage capacity are non-trivial; it is predicted that by 2025, between 100 million and 2 billion human genomes could have been sequenced. This volume of data could require up to 40 exabytes of storage [100] and will require significant investment to manage [101,102]). The conceptual challenges of how we collect, analyse and treat the resulting insights gleaned from this data are key issues that must be addressed to ensure any work with big data is valid. These considerations are discussed below. As well as conventional concerns about data quality [103] 'big data' pose unique challenges as they are not, unlike traditional scientific data, representationally sampled and are widely collected without prior hypothesis. Data being harvested without a clear purpose from the outset can lead to unidentified systematic biases [104] or problems down-the-line as data collection is updated to use new, improved schematics resulting in disparate datasets [105]. At a deeper level, if we blindly rely on insight from data with no context or where underlying mechanisms are not well described, our results might easily be confounded in ways that are not detectable.

Some practitioners are concerned that large volumes of data with no context may replace domain knowledge and scientific rigor, with the risk that the process of research in areas such as drug discovery moves from a scientific to an engineering discipline [106].

The mathematical model is key to the analysis of any large dataset and any collected data is only as useful as the model that represents it. Context in data (such as clinical annotation) can be hard to interpret at scale and is even harder to maintain when data are reduced to fit into a model [99]. Another concern when choosing a suitable model is how well the mechanism of your analysis can be interpreted. In traditional scientific processes each step of an experiment will be well understood and results should be easily interpreted to help both validate results and hopefully better understand underlying mechanics of the observed process. In mathematical models this isn't always the case, a neural network for example embodies no model of decision or even a problem domain, it is effectively black box that provides predictions of future events based on unauditable processes [107] and should be used cautiously even as a complement to experimental validation.

Even without mistakes at the collection or modelling stage, 'big data' analysis lends itself to number of statistical problems such as high false error rates [108] and overfitting [99]. Sound statistical practices, such as ensuring high-quality data, incorporating sound domain knowledge, and developing an overall strategy for modelling and validating problems, are even more crucial in big data analysis than it has been traditionally [109]. The importance of this statistical rigor is highlighted in well publicised cases such Potti et al. [70] where uncorrected sources of variation and inappropriate statistical methods led to cancelled clinic trials and a full retraction.

Many argue that big data should only ever be used as a complement to experimental validation and indeed there is currently much discussion within the Cancer Target Discovery and Development Network as to what level of experimental evidence is required to complement insights extracted from big data analysis [110].

5.2 Data standards

Perhaps the result of being an emerging field many of the categories of biological data captured do not have a conventional set of standards and as such many study reports and even repositories provide data with disparate format and labelling schemes [111]. Although the problem is computationally tractable, a lack of coherent standards can result in some difficulties when attempting to integrate different data types. As the field matures and certain standards are adopted over others we expect to see a continued consolidation of these standards [112].

Although improved standards will be a welcome improvement, the pace at which 'omic' techniques are progressing may still leave behind legacy data with redundant formats and incomparable datasets. To take just one example the ease and cost of generating gene expression data has been much improved with the advent of RNA sequencing. Yet even this single technique requires a number of choices to be made - from enrichment of mRNA in the laboratory to decisions about whether to map reads to the genome or to the annotated transcriptome [113]. Projects such as ICGCmed explicitly aim to enable comparability of results from experiments across the world and to build in some future proofing. Inevitably any static set of standards will be and should be overtaken by new innovations. In the author's view, the best that can be hoped for is a standards committee to agree, and promote best practice within the consortia; clear documentation showing the end user what analyses have

been carried out; a planned programme of revisions to standards; and finally a series of experiments carried out using both old and new standards in order to validate changes and throw light on the extent of comparability.

Even with potentially improved standards data in biological data repositories are not always easy to access. The Genomic Data Commons (GDC), one of the richest sources of data on the molecular biology of cancer, is also one of the hardest to navigate. Any substantial download requires an ability to use the terminal and write short scripts and it is not always straightforward to associate the files with the samples from which they came. For the GDC to realise its aim as a foundation of future cancer research, enabling remote experiments via cloud-based technologies the GDC will need to ensure that documentation about experimental technologies, pre-processing pipelines, and database terms is comprehensive and easy to find, and that there is an improved focus on the user experience of downloading data and matching samples to data. Despite there being copious guidance on the website, there is very little to encourage the novice user. Providing online tutorials and workshops would be a good way to introduce new users to these resources. Although the raw data may be difficult to manipulate, many of the online tools (e.g. CBioPortal [38]) allow cancer researchers to easily manipulate and explore previously pre-processed data.

Biological big data offers an unprecedentedly resource for cancer drug discovery but with this opportunity comes the risk of misapplication of the data. As the volume of data available for study increases so does the likelihood of false error or overfitting [108] unless we employ rigorous statistical fail-safes [114]. We must also ensure that any models used for analysis are appropriate for the dataset or risk losing important signal as the data is fitted to the model [99]. Even with best practice use of data in

place cancer biology is too complex to rely on data blindly and as such big data analyses should only ever be used as a complement to experimental validation [110].

Another notable weakness in contemporary genomic ‘big data’ is the sparsity of paired clinic data such as cancer diagnostic data [115], response to treatment regimes and patient lifestyle data. Fortunately, ICGCmed, the next phase of the ICGC project, aims to link current (and new) multi-platform data to clinical and health information [116] with discussion of making the inclusion of clinical data a requirement for data submission, this stipulation would certainly incentivise researchers to provide a previously relatively neglected data of great value to drug discovery research groups. This gives rise to ethical issues that are similar to those encountered in traditional medical genetics. The large amount of information that could be obtained by ‘Big Data’ approaches requires a careful evaluation on how to implement core ethical principles including: informed consent, privacy and data ownership and sharing, technology regulation, and the issues of access to the data [120]. In the case of the ICGC, concerns over ethical or privacy of the data are addressed through ICGCs stringent policies on restricted data access.

6 Conclusions

Continued advances in technology has enabled us to analyse increasingly large datasets and international initiatives to curate and collate this data provide an unprecedented resource for contemporary research communities. There are however still improvements to be made. Issues regarding collection, quality control, standardisation, access and statistical rigour that will need to be addressed before these methods and resources are fully matured.

Cancer chemotherapy is currently undergoing a step change, transitioning from traditional, general treatments such as potentially harmful cytotoxic agents to a much more selective approach including the use of targeted and even tailored therapies. This development is in large part a result of the growing availability of biological ‘big data’ across a wide range of platforms and tools developed to transform that data into actionable insight.

The need for improved cancer therapies has led to a wide range of pan-cancer studies that integrate large volumes of data from many platforms. These studies have allowed us to build both our knowledge of the mechanics behind the cancers and the resulting weaknesses that can potentially be exploited as drug targets. Further studies have provided ways for us to predict the suitability of these drug targets potentially saving some of the significant time and cost that come as a result of failure in the drug discovery process.

Although ‘big’ data cannot be used as any other than a complement to experimental validation it continues to prove a crucial part of modern drug discovery as we map out the landscapes of cancers and better characterise different cancer subtypes to ultimately provide a more effective, resilient and targeted set of cancer therapies.

7 Expert opinion

The national and international coordination of ‘big data’ multi-omic approaches to characterise tumours has produced an unparalleled opportunity for the cancer community to systematically identify all cancer drivers. The continued growth in volume and availability of genetic data presents many opportunities in the field of cancer drug discovery. These improvements will include identifying markers for early

detection of disease, more specific criteria and methods for diagnoses and prognoses, and interventions based on matching the patient's disease molecular subtype with the most effective combinations of therapies. However improvements are required in both in capturing and storing the data and in its ability to be accessed and analysed by informaticians and cancer researchers alike.

Next generation sequencing and other 'omic' technologies are enabling better stratification of a wide range of cancers, which will eventually lead to targeted therapies and personalised medicines. The best known example of patient stratification is the analysis of biomarkers for patients with breast cancer that are used to determine treatment regimes. Tests and endocrine therapies already exist for patients testing positive for elevated levels of HER2, estrogen or progesterone and these complement chemotherapy, radiation therapy and surgery, the current standard-of-care in cancer treatment. One of the most pressing needs in cancer bioinformatics is to provide this type of characterisation for other cancers, identifying stable patient cohorts with similar therapeutic needs and potential outcomes and reducing the use of aggressive therapies where they are not warranted.

Such approaches offer the possibilities of developing biomarkers to improve existing therapies, providing information for potential drug design, or simply improve the accuracy of prognoses and have been used to stratify a number of cancers including breast cancers, ovarian cancer [117,118] and Acute Myeloid Leukaemia [119].

Predicting and overcoming resistance is also a major challenge for targeted cancer therapies as it was for the earlier chemotherapies. Some of the more recent targeted therapies have provided at best only short-lived remissions before resistance takes hold. This is due either to the prior existence of resistant sub-clones or continued evolution of the tumour under the selective pressure of a drug regime [8,9]. The goal

of using a tailored approach to treat all tumour patients may therefore require combination therapies with one or more drugs to prevent resistance developing.

‘Big data” can also facilitate the success of clinical trials. Testing novel compounds against characterised cell-line collections may help enable preclinical stratification to identify sensitive cancers. Data produced from these types of study can be used to inform clinical trial design and aid the development of personalised therapeutic regimens

The continued growth in volume and availability of ‘big data’ presents many opportunities in the field of cancer drug discovery and offers promising ways to inform decisions at each step of the discovery pipeline from drug target identification and corresponding molecule selection to stratifying patients allowing for better targeted trials and reduced attrition rates. Using large-scale integrated data, researchers, scientists, policymakers and clinicians will be able to work with patients and healthcare providers to deliver truly personalised cancer care.

8 References

[1] Varmus H, Kumar HS. Addressing the Growing International Challenge of Cancer: A Multinational Perspective. Sci. Transl. Med. 2013;5:175cm2–cm175cm2.

[2] Pearl LH, Schierz AC, Ward SE, et al. Therapeutic opportunities within the DNA damage response. Nat. Rev. Cancer. 2015;15:166–180.

[3] Varmus H. The New Era in Cancer Research. Science. 2006;312:1162–1165.

[4] Mukherjee S. The Emperor of All Maladies: A Biography of Cancer. Simon and Schuster; 2011.

[5] Yap TA, Workman P. Exploiting the Cancer Genome: Strategies for the Discovery and Clinical Development of Targeted Molecular Therapeutics. Annu. Rev. Pharmacol. Toxicol. 2012;52:549–573.

[6] Garnock-Jones KP, Keating GM, Scott LJ. Trastuzumab: A review of its use as adjuvant treatment in human epidermal growth factor receptor 2 (HER2)-positive early breast cancer. Drugs. 2010;70:215–239.

[7] Stagno F, Stella S, Spitaleri A, et al. Imatinib mesylate in chronic myeloid leukemia: frontline treatment and long-term outcomes. Expert Rev. Anticancer Ther. 2016;16:273–278.

[8] Garraway LA, Jänne PA. Circumventing Cancer Drug Resistance in the Era of Personalized Medicine. Cancer Discov. 2012;2:214–226.

[9] Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. Nat. Biotechnol. 2012;30:679–692.

[10] Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153:17–37.

[11] Stratton MR. Exploring the Genomes of Cancer Cells: Progress and Promise. Science. 2011;331:1553–1558.

[12] Ciriello G, Sinha R, Hoadley KA, et al. The molecular diversity of Luminal A breast tumors. Breast Cancer Res. Treat. 2013;141:409–420.

[13] Brennan CW, Verhaak RGW, McKenna A, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155:462–477.

*** Multi-platform analysis of over 500 glioblastoma tumours to identify cancer drivers and suggest therapeutic strategies.**

[14] Pereira B, Chin S-F, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat. Commun. 2016;7:11479.

*** Multi-platform analysis of breast cancers.**

[15] Giannakis M, Mu XJ, Shukla SA, et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Rep. 2016;17:1206.

[16] Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science. 2013;339:1546–1558.

**** Analysis of mutations from 26 types of cancer to identify drivers. The paper then highlights possible therapeutic strategies**

[17] Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153–158.

**** Mutational analysis of the kinome from 210 human cancers to identify cancer drivers.**

[18] Wood LD, Parsons DW, Jones S, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318:1108–1113.

**** Following the exome sequencing of 11 breast and 11 colorectal cancers,**

the authors describe their cancer genomic landscapes. Cancers are composed of “a handful of commonly mutated gene "mountains" and a much larger number of gene "hills" that are mutated at low frequency”.

[19] Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–339.

[20] Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci. Rep. 2013;3:2650.

*** Analysis of mutational data from 12 cancer studies and identify 291 high-confidence cancer driver genes.**

[21] Davoli T, Xu AW, Mengwasser KE, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155:948–962.

[22] Yoon S-H, Kim J-S, Song H-H. Statistical inference methods for detecting altered gene associations. Genome Inform. 2003;14:54–63.

[23] Pickering CR, Zhang J, Yoo SY, et al. Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. Cancer Discov. 2013;3:770–781.

*** Multi-platform analysis of oral squamous cell carcinoma.**

[24] Zheng S, Cherniack AD, Dewal N, et al. Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016;30:363.

*** Multi-platform characterisation of adrenocortical carcinoma**

[25] Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502:333–339.

[26] Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. U. S. A. 2013;110:4245–4250.

[27] Workman P, Al-Lazikani B. Drugging cancer genomes. Nat. Rev. Drug Discov. 2013;12:889–890.

[28] Rubio-Perez C, Tamborero D, Schroeder MP, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell. 2015;27:382–396.

*** Paper describes drug repositioning opportunities following the analysis of multi-platform data from 28 tumour types.**

[29] Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat. Rev. Drug Discov. 2016;16:19–34.

*** The authors map 1,578 FDA-approved drugs to their therapeutic**

indication. They show that 70% of small molecule drugs act through 'privileged' protein functional families (GPCRs, ion channels, kinases and nuclear receptors), and highlight that only 5% of identified cancer driver genes are targeted by current cancer therapies.

[30] Thatcher N, Chang A, Parikh P, et al. Gefitinib plus best supportive care in previously treated patients with refractory advanced non-small-cell lung cancer: results from a randomised, placebo-controlled, multicentre study (Iressa Survival Evaluation in Lung Cancer). Lancet. 2005;366:1527–1537.

[31] Shepherd FA, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N. Engl. J. Med. 2005;353:123–132.

[32] Stinchcombe TE, Socinski MA. Gefitinib in advanced non-small cell lung cancer: does it deserve a second chance? Oncologist. 2008;13:933–944.

[33] Lindeman NI, Cagle PT, Beasley MB, et al. Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. J. Mol. Diagn. 2013;15:415–453.

[34] Tokheim C, Papadopoulis N, Kinzler KW, et al. Evaluating the Evaluation of Cancer Driver Genes [Internet]. 2016. Available from: http://dx.doi.org/10.1101/060426.

[35] Baeissa HM, Benstead-Hume G, Richardson CJ, et al. Mutational patterns in oncogenes and tumour suppressors. Biochem. Soc. Trans. 2016;44:925–931.

[36] Baeissa H, Benstead-Hume G, Richardson CJ, et al. Identification and analysis of mutational hotspots in oncogenes and tumour suppressors. Oncotarget [Internet]. 2017; Available from: http://dx.doi.org/10.18632/oncotarget.15514.

[37] Luo J, Solimini NL, Elledge SJ. Principles of cancer therapy: oncogene and non-oncogene addiction. Cell. 2009;136:823–837.

*** Review describing non-oncogene addiction and its therapeutic opportunities.**

[38] Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. 2013;6:l1.

[39] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 2005;102:15545–15550.

[40] Huang DW, Sherman BT, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35:W169–W175.

[41] Reimand J, Kull M, Peterson H, et al. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res.

2007;35:W193–W200.

[42] Gundem G, Lopez-Bigas N. Sample-level enrichment analysis unravels shared stress phenotypes among multiple cancer types. Genome Med. 2012;4:28.

[43] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.

[44] Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38:W214–W220.

[45] Hofree M, Shen JP, Carter H, et al. Network-based stratification of tumor mutations. Nat. Methods. 2013;10:1108–1115.

[46] Khoo KH, Verma CS, Lane DP. Drugging the p53 pathway: understanding the route to clinical efficacy. Nat. Rev. Drug Discov. 2014;13:314–314.

* **Therapeutic strategies that target the p53 pathway.**

[47] Hartwell LH, Szankasi P, Roberts CJ, et al. Integrating genetic approaches into the discovery of anticancer drugs. Science. 1997;278:1064–1068.

[48] Megchelenbrink W, Katzir R, Lu X, et al. Synthetic dosage lethality in the human metabolic network is highly predictive of tumor growth and cancer patient survival. Proc. Natl. Acad. Sci. U. S. A. 2015;112:12217–12222.

[49] Kaelin WG Jr. The concept of synthetic lethality in the context of anticancer therapy. Nat. Rev. Cancer. 2005;5:689–698.

[50] Bryant HE, Schultz N, Thomas HD, et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature. 2005;434:913–917.

[51] Brown JS, Kaye SB, Yap TA. PARP inhibitors: the race is on. Br. J. Cancer. 2016;114:713–715.

[52] Dixon SJ, Fedyshyn Y, Koh JLY, et al. Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 2008;105:16653–16658.

[53] Dixon SJ, Andrews B, Boone C. Exploring the conservation of synthetic lethal genetic interaction networks. Commun. Integr. Biol. 2009;2:78–81.

[54] Iorns E, Lord CJ, Turner N, et al. Utilizing RNA interference to enhance cancer drug discovery. Nat. Rev. Drug Discov. 2007;6:556–568.

* **Review describing RNAi approaches in cancer drug discovery**

[55] Madhukar NS, Elemento O, Pandey G. Prediction of Genetic Interactions Using Machine Learning and Network Properties. Front Bioeng Biotechnol.

2015;3:172.

[56] Jerby-Arnon L, Pfetzer N, Waldman YY, et al. Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. Cell. 2014;158:1199–1209.

[57] Wang X, Simon R. Identification of potential synthetic lethal genes to p53 using a computational biology approach. BMC Med. Genomics. 2013;6:30.

[58] Zhang F, Fan Z, Min W, et al. Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. J. Bioinform. Comput. Biol. 2015;13:1541002.

[59] Chatr-Aryamontri A, Breitkreutz B-J, Oughtred R, et al. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43:D470–D478.

[60] Chipman KC, Singh AK. Predicting genetic interactions with random walks on biological networks. BMC Bioinformatics. 2009;10:17.

[61] Paladugu SR, Zhao S, Ray A, et al. Mining protein networks for synthetic genetic interactions. BMC Bioinformatics. 2008;9:426.

[62] Pandey G, Zhang B, Chang AN, et al. An integrative multi-network and multi-classifier approach to predict genetic interactions. PLoS Comput. Biol. [Internet]. 2010;6. Available from: http://dx.doi.org/10.1371/journal.pcbi.1000928.

[63] Jacunski A, Dixon SJ, Tatonetti NP. Connectivity Homology Enables Inter-Species Network Models of Synthetic Lethality. PLoS Comput. Biol. 2015;11:e1004506.

[64] Wu M, Li X, Zhang F, et al. In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. Cancer Inform. 2014;13:71–80.

[65] Workman P, Al-Lazikani B, Clarke PA. Genome-based cancer therapeutics: targets, kinase drug resistance and future strategies for precision oncology. Curr. Opin. Pharmacol. 2013;13:486–496.

[66] Forbes SA, Beare D, Bindal N, et al. COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. Curr. Protoc. Hum. Genet. 2016;91:10.11.1–10.11.37.

*** Excellent resource that documents somatic cancer mutations.**

[67] Scholl C, Fröhling S, Dunn IF, et al. Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. Cell. 2009;137:821–834.

[68] Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462:108–112.

[69] Wang Y, Ngo VN, Marani M, et al. Critical role for transcriptional repressor Snail2 in transformation by oncogenic RAS in colorectal carcinoma cells.

Oncogene. 2010;29:4658–4670.

[70] Potti A, Dressman HK, Bild A, et al. Retraction: Genomic signatures to guide the use of chemotherapeutics. Nat. Med. 2011;17:135–135.

[71] Costa-Cabral S, Brough R, Konde A, et al. CDK1 Is a Synthetic Lethal Target for KRAS Mutant Tumours. PLoS One. 2016;11:e0149099.

[72] Vicent S, Chen R, Sayles LC, et al. Wilms tumor 1 (WT1) regulates KRAS-driven oncogenesis and senescence in mouse and human models. J. Clin. Invest. 2010;120:3940–3952.

[73] Helleday T. Cancer phenotypic lethality, exemplified by the non-essential MTH1 enzyme being required for cancer survival. Ann. Oncol. 2014;25:1253–1255.

[74] Gad H, Koolmeister T, Jemth A-S, et al. MTH1 inhibition eradicates cancer by preventing sanitation of the dNTP pool. Nature. 2014;508:215–221.

[75] Campbell J, Ryan CJ, Brough R, et al. Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. Cell Rep. 2016;14:2490–2501.

[76] Aromataris E, Fernandez R, Godfrey CM, et al. Summarizing systematic reviews: methodological development, conduct and reporting of an umbrella review approach. Int. J. Evid. Based Healthc. 2015;13:132–140.

[77] Cunanan KM, Iasonos A, Shen R, et al. An efficient basket trial design. Stat. Med. [Internet]. 2017; Available from: http://dx.doi.org/10.1002/sim.7227.

[78] Redig AJ, Jänne PA. Basket Trials and the Evolution of Clinical Trial Design in an Era of Genomic Medicine. J. Clin. Oncol. 2015;33:975–977.

[79] Renfro LA, Sargent DJ. Statistical controversies in clinical research: basket trials, umbrella trials, and other master protocols: a review and examples. Ann. Oncol. 2016;mdw413.

[80] Hopkins AL, Groom CR. The druggable genome. Nat. Rev. Drug Discov. 2002;1:727–730.

*** Original paper describing the set of possible human drug targets.**

[81] Jamali AA, Ferdousi R, Razzaghi S, et al. DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discov. Today. 2016;21:718–724.

[82] Meyers J, Brown N, Blagg J. Mapping the 3D structures of small molecule binding sites. J. Cheminform. [Internet]. 2016;8. Available from: http://dx.doi.org/10.1186/s13321-016-0180-0.

[83] Kumari P, Nath A, Chaube R. Identification of human drug targets using machine-learning algorithms. Comput. Biol. Med. 2015;56:175–181.

[84] Patel MN, Halling-Brown MD, Tym JE, et al. Objective assessment of cancer

genes for drug discovery. Nat. Rev. Drug Discov. 2013;12:35–50.

[85] Richardson CJ, Gao Q, Mitsopoulous C, et al. MoKCa database--mutations of kinases in cancer. Nucleic Acids Res. 2009;37:D824–D831.

[86] Printz C. Genomic Data Commons ushers in new era for information sharing. Cancer. 2016;122:2777–2778.

[87] Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. 2015;19:A68–A77.

[88] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 2013;45:1113–1120.

[89] Zhang J, Baran J, Cros A, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database . 2011;2011:bar026.

[90] Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2:401–404.

[91] Sakoparnig T, Fried P, Beerenwinkel N. Identification of constrained cancer driver genes based on mutation timing. PLoS Comput. Biol. 2015;11:e1004027.

[92] Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45:D945–D954.

[93] Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–607.

[94] Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–D961.

[95] Liu C, Bai B, Skogerbø G, et al. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res. 2005;33:D112–D115.

[96] Amaral PP, Clark MB, Gascoigne DK, et al. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011;39:D146–D151.

[97] Cook CE, Bergman MT, Finn RD, et al. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 2016;44:D20–D26.

[98] Jacobs A. The Pathologies of Big Data. Queueing Syst. 2009;7:10.

[99] Boyd D, Crawford K. CRITICAL QUESTIONS FOR BIG DATA. Inf. Commun. Soc. 2012;15:662–679.

[100] Stephens ZD, Lee SY, Faghri F, et al. Big Data: Astronomical or Genomical?

PLoS Biol. 2015;13:e1002195.

[101] Lynch C. Big data: How do your data grow? Nature. 2008;455:28–29.

[102] Trelles O, Prins P, Snir M, et al. Big data, but are we ready? Nat. Rev. Genet. 2011;12:224.

[103] Weng C, Kahn MG. Clinical Research Informatics for Big Data and Precision Medicine. Yearb. Med. Inform. 2016;211–218.

[104] Fallik D. For Big Data, Big Questions Remain. Health Aff. 2014;33:1111–1114.

[105] Lazer D, Kennedy R, King G, et al. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014;343:1203–1205.

[106] Callebaut W. Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences. 2012;43:69–80.

[107] Clarke R. Big data, big risks. Information Systems Journal. 2015;26:77–90.

[108] Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2:e124.

[109] Hoerl RW, Snee RD, De Veaux RD. Applying statistical thinking to "Big Data" problems. Wiley Interdiscip. Rev. Comput. Stat. 2014;6:222–232.

[110] Cancer Target Discovery and Development Network. Transforming Big Data into Cancer-Relevant Insight: An Initial, Multi-Tier Approach to Assess Reproducibility and Relevance. Mol. Cancer Res. 2016;14:675–682.

[111] Elefsinioti A, Bellaire T, Wang A, et al. Key factors for successful data integration in biomarker research. Nat. Rev. Drug Discov. 2016;15:369–370.

[112] Rocca-Serra P, Salek RM, Arita M, et al. Data standards can boost metabolomics research, and if there is a will, there is a way. Metabolomics [Internet]. 2015;12. Available from: http://dx.doi.org/10.1007/s11306-015-0879-3.

[113] Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

[114] Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1:140216.

[115] Al-Shahrour F, Malats N, Valencia A, et al. CANCEROMATICS III - Tumor Heterogeneity [Internet]. 2016. Available from: https://www.cnio.es/eventos/index.asp?ev=2&cev=138.

[116] Jennings JL, Hudson TJ. Abstract 130: International Cancer Genome Consortium (ICGC). Cancer Res. 2016;76:130–130.

[117] Wang C, Machiraju R, Huang K. Breast cancer patient stratification using a

molecular regularized consensus clustering method. Methods. 2014;67:304–312.

[118] Riester M, Wei W, Waldron L, et al. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. J. Natl. Cancer Inst. [Internet]. 2014;106. Available from: http://dx.doi.org/10.1093/jnci/dju048.

[119] Kelly AD, Heike K, Jumpei Y, et al. Abstract B22: Genome-wide methylation analysis reveals an independently validated CpG island methylator phenotype associated with favorable prognosis in acute myeloid leukemia. Clin. Cancer Res. 2015;21:B22–B22.

[120] Allain DC, Ormond KE. Ethical Issues in Genetic and Genomic Research. Ethical Dilemmas in Genetics and Genetic Counseling. 2014. p. 186–208.

Article Highlight Box

- Large-scale studies of genetic mutations have identified at best a very small number of highly recurrent altered genes, with increasingly long tails of much more infrequently altered genes. Studies that cluster these rarer mutations are providing insight into biological understanding of the pathways involved in cancers.
- The patterns of mutations within driver genes enable identification of genes into oncogenes and tumour suppressors. Tumour suppressors cannot often be directly targeted. Instead analysis is required to find gene products that are synthetically lethal to the tumour suppressor and can be inhibited pharmacologically. The DNA damage response pathways are particularly fruitful sources of tumour suppressors that can be targeted in this way.
- The genetic dependencies of cancer cell lines can be profiled, identifying where the cancer has become addicted to support from altered pathways, allowing new therapeutic options. Analysis of omic data from cell lines tested with novel compounds has allowed genetic, lineage, and gene-expression-based predictors of drug sensitivity.

- Proteins can be assessed for druggability by modelling 3D structure and analysing the extent to which ‘pockets’ in the protein bind pharmacologically suitable molecules with high affinity and specificity. Where proteins are not suitable analysis is needed to find synthetic dosage lethal partners.
- Much of the data produced from these large-scale studies is publicly available, together with tools allowing integrated analysis.